DOCUMENT RESUME

ED 417 599 FL 025 121

AUTHOR Nakamura, Yuji
TITLE Involving Factors of Fairness in Language Testing.
PUB DATE 1997-09-25
NOTE 21p.; Based on a paper presented at the Annual Meeting of the Language Testing Research Colloquium (19th, Orlando, FL, October 6- 9, 1997).
AVAILABLE FROM The "Journal of Communication Studies" is published by Keizai University, Tokyo, Japan.
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
JOURNAL CIT Journal of Communication Studies; n7 p3-21 Sep 1997
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Behavior Patterns; Comparative Analysis; *English (Second Language); Interrater Reliability; *Interviews; Language Laboratories; *Language Tests; Rating Scales; Second Language Instruction; Student Attitudes; Surveys; *Test Bias; Test Format; Test Items; *Testing; *Verbal Tests

ABSTRACT

This study investigated the effects of three aspects of language testing (test task, familiarity with an interviewer, and test method) on both tester and tested. Data were drawn from several previous studies by the researcher. Concerning test task, data were analyzed for the type of topic students wanted most to talk about or preferred not to talk about, and whether they had similar preferences for Japanese and English tests. Concerning the interviewer factor, data were analyzed for whether the interviewer was a classroom teacher, whether teacher and interviewer could share a common conversation topic, and whether the interviewers were interested in topics the students respond to. Student preferences for oral test method, direct or semi-direct and type of interaction used to elicit speech, were also analyzed. Results indicate that at different proficiency levels, students perform differently on direct and semi-direct tests, and interviewers' choice of test questions influenced student performances and may have even influenced raters' ratings. Implications for fairness in testing are considered. Contains 18 references. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 417 599

Involving Factors of Fairness in Language Testing

Yuji Nakamura

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Yuji Nakamura

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

- This document has been reproduced as received from the person or organization originating it.
- ☐ Minor changes have been made to improve reproduction quality.
- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

コミュニケーション科学第7号抜刷 (1997年9月)
東京経済大学

FL951a1

BEST COPY AVAILABLE

Involving Factors of Fairness in Language Testing

Yuji Nakamura

I. Theoretical Background

To what degree, as a teacher or tester, can we give a fair test to students? What can or should teachers do to give a fairer test to students? Some students are good at interview tests, while others are skillful in tape-mediated speaking tests. Some tend not to speak about families, whereas others are likely to. Still some students are given lower grades just because they are less skillful in oral summary even if they can comprehend the reading material. There seems to be many problems to be solved in the issue of fairness.

Fairness can be defined as the degree to which a test treats every student the same or the degree to which it is impartial (Brown 1996). We, as test users or language teachers, should do our best so that our personal feelings do not interfere with fair assessment of the students or bias the assignment of scores. Our target is to give each student an equal chance to do well on the test. Therefore, we often do everything in our power to find test questions, administration procedures, scoring methods, and reporting policies that optimize the chances that each student will receive equal and fair treatment. (cf. Brown 1996).

There are a number of elements related to the test fairness such as test method, test task, rater judgment, test taker characteristics, and cultural background. There is a claim that two versions of the same test cannot be strictly comparable. Accordingly, a method difference between two tests (an interview test versus a tape-mediated test; group speaking test versus individual speaking test) produces a method effect (Shohamy 1996). Another claim is about the conflict between fairness and face validity. (Lumley 1996).

II. Purpose of the research

In this paper, fairness will be investigated mainly from three viewpoints:

1) test task 2) familiarity with an interviewer 3) test method. The eventual goal of this research is to show how these factors influence the testers and testees, and to help us give a much fairer test to students.

III. Research design and methods

This research is conducted in terms of three aspects: test task, familiarity with an interviewer and test method. Several questions are asked in order to obtain enough information in individual aspects.

In the task factor the question is : What topic do students most want to talk about? or what topic do they not want to speak about? Also, they are asked whether they ask the same type of questions or complaints in Japanese which are used in the English test.

In the interviewer/rater factor, the following questions are asked: 1) Is the interviewer a classroom teacher or not? 2) Can the teacher or interviewer share the common topic for conversation with students? 3) Are the interviewers interested in the topics the students can easily respond?

In the method factor the questions are: 1) Even in the speaking test, a direct test is favored by some students, while a semi-direct test is welcomed by others. Since the test method effect is influential, how can we make the trade off between a fair judgment of speaking and the test method preference? 2) In a speaking test, the test methods such as a monologue type test, a dialogue type test, a discussion type test, a debate type test, are influential factors in getting the students to speak out and give a fair judgment of their speaking ability. Since the test method effect is unavoidable, how can we take it into consideration for the assessment?

IV. Procedures

1. Test Task Aspect

1.1. Test results of Nakamura (1993) were used to obtain information on the students' preferences for speech making topics (n=80).

1.2. Test results of Nakamura (1993) were again used to obtain information on the students' preferences as a visual material description topic (n=80).

1.3. Questionnaire results (1997) were used to obtain another piece of information on the students' preferences as a speech making topic (n=73).

Question used was: What topic do you want to talk about in a speech making test?

1.4. Questionnaire results (1997) were used to get the students' answering patterns or behavioral patterns in English and in Japanese (n=40).

Question used was : In the following nine individual contexts, would you say or act the same way in English and in Japanese as in the assigned context? Why ?

2. Interviewer/Rater Aspect

Research results of Nakamura (1996) were used to obtain information on the interviewer/rater scoring tendency.

3. Test Method Aspect

3.1. Questionnaire results (1997) were used to obtain information on the students' preferences about the test method (either a direct face to face interview test with a native speaker of English or a semi-direct speaking test in the language laboratory) (n=44).

Question used was: Which do you prefer, an interview test or a language laboratory speaking test? Why ?

3.2. Research results of Nakamura (1996) were used to show the results of the factor analysis of the three test methods (an interview test, a semi-direct speaking test, and a writing test) (n=80).

3.3. Research results of Nakamura (1997) were used to demonstrate the relationship of the four test methods (a semi-direct speaking test, an interview test, a writing test, and a group discussion test) (n=32).

V. Results and Discussion

1. Examination of Test Task or Test Topic by focusing on the students or test takers

As the item analysis is conducted by analyzing each item in terms of item difficulty, item discrimination and the function of the item, we should analyze the task item by investigating how the test topic has been chosen. In other words, what topics or topic areas are students most likely to choose in a speaking test situation? First, we will look at the results from the preferences conducted in the real speaking tests.

Table 1

frequencies of task item for making a speech

item	freq.	topic
1	9	my friends
2	20	my family
3	16	part-time work
4	17	my hobbies
5	15	traveling
6	0	fashion
7	1	telephone conversation
8	2	college life

n=80

Table 1 demonstrates that in the real speech making tests the topics (my friends, my family, part-time work, my hobbies, traveling) are common among students as a topic for making a speech, and among them My Family is the most popular one. On the other hand, topics such as fashion, telephone conversation and college life are less common.

Let us examine the results from the questionnaire research about their preferences for the topic of the speech.

Table 2

topics students want to talk about

item	freq.	topic
1	10	sports
2	13	family
3	9	friends
4	13	hobbies
5	3	part-time job
6	1	traveling
7	1	food
8	4	college life
9	4	dreams
10	3	home town
11	2	music
12	2	current events
13	3	club activities
14	1	internet
15	1	news
16	1	movies
17	1	Japan
18	1	fashion

n=73

Table 2 indicates that, in their mind, students have various topics for the speech making even if the frequencies are not equal. In Table 2, the topics listed by the students as ones they want to talk about are hobbies, sports, family, friends, college life, dreams, part-time job, traveling, food, home town, music, current events, club activities, internet, news, movies, Japan, and fashion.

From these two types of data, it might be said there is a mismatch between the topic teachers choose and that students want to talk about with much confidence.

Let us look at the results of the actual test of the visual material description.

Table 3

fequencies of topic item for visual material description

item	freq.	topic
1	5	picture(pic) of Superman
2	5	pic of Marilyn Monroe
3	0	pic of a boxer
4	0	pic of a singer
5	2	pic of late Japanese Emperor
6	2	pic of Shakespeare
7	9	pic of E.T.
8	5	pic of late Kennedy
9	0	pic of a lady
10	0	pic of Gorbachev
11	0	a description of mailing system
12	6	a map of the U.S.
13	12	a cartoon of Fuji Santaro
14	14	a cartoon of Fuji Santaro
15	1	a graph of employment rate
16	5	an advertisement of a language school
17	5	a TV program
18	0	a train schedule
19	2	a brochure of a city tour
20	1	a seminar schedule
21	6	an itinerary for Dr. Brown

n=80

Table 3 suggests that among 21 choices, cartoons are the most frequently chosen and for the students cartoons are easy to produce sentences. This is an interesting finding. Some of the popular topics are the pictures of the things or people that students are familiar

with, which is not surprising. A TV program or itinerary which has some self-explanatory information in the material is also common. On the other hand, the pictures that are not familiar to students may not be chosen so much or even at all. Although teachers do not need to use the choices which are not statistically significant, they always try to make the trade off between the practical test method and the best fit as a test topic.

Lastly, we will look at the influence of the native language (culture) on the behavior in the target (foreign) language. The questionnaire consisting of nine items was conducted in the following way as shown in Table 4.

Table 4

Questionnaire

1) In the following nine contexts, would do you do so or would you say so in Japanese in the Japanese context? (Yes:JYes, No:JNo)
2) In the following nine contexts, would do you do so or would you say so in English in the English context? (Yes:EYes, No:ENo)

Comments for each item

Item 1 (Context:Apologizing and making an excuse)

You are late for your class. You missed the school bus. Please apologize and make an excuse to your teacher.

1) JENo

I go to my seat without making an excuse.

It sounds like I am making an excuse.(6)

I don't want to disturb the class.(2)

2) JNo

It is my fault to be late.(3)

It is not usual to make an excuse in a Japanese class.(2)

It sounds like I am making an excuse.(2)

It is troublesome to make an excuse.

It is highly regarded not to make an excuse in Japan.

Teachers won't listen to students' excuses.

3) ENo

My English is not efficient enough to make an excuse.

Item 2 (Context:Complaining and requesting)

You are in a non-smoking section of a waiting room at the airport. Someone started smoking. You have a cold or a sore throat. Please complain about it and request him/her to stop it.

1) JENo

I do not want to be bothered by the person.

I am afraid that would give offense.

I will be patient with it.
I will express my feeling by my attitude.
2) JNo
I will wait until someone else complains about it.
I will endure it.
I myself stay away from the person. It is embarrassing to make a complaint.
I will keep silent. I do not want to cause trouble.(2)
3) ENo
I am not sure whether I can make myself understood in this situation.(2)
I cannot refute the person if he or she argues back.(3)

Item 3 (Context: Asking for repetition)
You didn't understand what your teacher said. You want the teacher to repeat it. Please make a request to your teacher.
1) JENo
I don't want to stop the class.(3)
I will wait until someone else asks about it.
I will ask my friend about it later.(4)
2) JNo
It is not usual to say so in Japan.
I am concerned about the reaction of the other students.(2)
3) ENo
I am afraid I have to say so too many times.
In English I will take a passive attitude.
I am not sure whether I can understand the second time.

Item 4 (Context: Greeting)
You happen to meet your high school teacher after a long interval. Please greet him.
1) JENo
It is troublesome to talk.
2) ENo
I am afraid I will only be answering the questions, not asking the questions.
I cannot explain my daily life in English.(2)
I am not accustomed to such a situation and I feel embarrassed.

Item 5 (Context: Disagreeing)
Your friend says jogging is a healthy activity. You don't agree with her. What do you say to her?
1) JNo
I don't want to devastate the atmosphere.
It is troublesome and useless to argue.
2) ENo
I don't want to be retuted.
I don't like argument.(5)
I am not confident on making a counter argument.

BEST COPY AVAILABLE

I tend to be persuaded without being aware of it in English.

Item 6 (Context: Interrupting)
Your supervisor is working in his office. You want to interrupt him for a moment to talk with him. What do you say?
1) JENo
I don't want to disturb the boss.
I will wait until the boss is finished with his work.
2) Jno
The boss might get angry if I interrupt his job.

Item 7 (Context: Warning)
Some children are playing baseball and almost break the window of your house. Please warn them.
1) JENo
I think children will understand the situation.
I will let the children play freely.
I will wait until the children recognize the danger.
If the children break the window, I will warn them.
I don't want to get involved with children.
Children will not seem to listen to me.(2)
It has nothing to do with me.
2) JNo
Giving a warning sounds harsh in Japanese.
3) ENo
I don't know how to warn them in English.(2)

Item 8 (Context: Offering)
You want to serve something to drink to a guest at your house. Please offer something to drink.
JENo
If they want to drink, they will drink by themselves.

Item 9(Context: Asking for information)
At a department store, please ask the receptionist where the stationery section is.
1) JENo
We can find it in the map.(2)
2) JNo
We know where it is in the Japanese context.
I will find it by myself.
3) ENo
I am afraid I cannot make myself understood in English.
I will try by myself and I will give it up if I cannot find the place.

Table 4 shows the raw data: comments obtained from the students. Before going into the detail of the data in Table 4, let us look at the statistical analysis of the results of the data in Table 5 .

Table 5

item 1

	JYes	JNo
EYes	10	19
ENo	2	9

df=1 χ^2=0.38 p>.05

item 2

	JYes	JNo
EYes	17	8
ENo	7	8

df=1 χ^2=1.77 p>.05

item 3

	JYes	JNo
EYes	21	7
ENo	3	9

df=1 χ^2=6.79 p<.01

item 4

	JYes	JNo
EYes	31	1
ENo	6	2

df=1 χ^2=1.82 p>.05

item 5

	JYes	JNo
EYes	25	5
ENo	4	6

df=1 χ^2=5.05 p<.05

item 6

	JYes	JNo
EYes	32	2
ENo	2	4

df=1 χ^2=10.39 p<.01

item 7

	JYes	JNo
EYes	20	4
ENo	2	14

df=1 χ^2=16.7 p<.01

item 8

	JYes	JNo
EYes	34	3
ENo	1	2

df=1 χ^2=4.16 p<.05

item 9

	JYes	JNo
EYes	25	6
ENo	5	4

df=1 χ^2=1.19 p>.05

N.B.
EYes: I would do so or I would say so in English in the English context.
Eno: I would not do so or I would not say so in English in the English context.
JYes: I would do so or I would say so in Japanese in the Japanese context.
JNo: I would not do so or I would not say so in Japanese in the Japanese context.

Table 5 shows that there is a statistically significant difference in the distribution in items (3,5,6,7,8), while there is no statistical difference in the distribution in items (1,2, 4,9). This result indicates the following things;

1) In items (3,5,6,7,8), there is a similar pattern in the native language context and in the target (English) language context about the students' behavioral activities. In other words, they would act or say in the English language context just the same way as in the Japanese language context.

2) In these items, there is influence from the language or cultural difference; therefore, if students make a mistake or cannot handle a problem in the English language context, which means that they lack English ability such as ability to offer in English, or ability to interrupt in English.

3) In the items (1,2,4,9), there is no statistically significant difference in their behavioral pattern. In other words, there is a variety in their behaving pattern in the English context.

Let us look at both Table 4 and Table 5, and analyze the patterns. One type may be the ones who choose "yes" in Japanese and "no" in English. One assumed explanation is that they know the standard code of expressing their opinions in these contexts so they are straightforward in Japanese. When they are not sure about the level of politeness in English, they would not say a word. Another type is the ones who choose "yes" in English but choose "no" in Japanese. There seems to be a cultural difference. In English they are more active and assertive , but they are more reserved and less assertive in Japanese. Therefore, it sometimes takes more time to speak out in English when the Japanese way of thinking disturbs their English utterances. Cultural differences can be noted in 2,4,5 and 9. In other words, the pattern of "yes" in Japanese and "no" in English varies.

In summary, raters should take into account some possibilities of students' background. What is behind their silence? What is behind their wrong answers? What is behind their behavior? There seems to be more involved in the students' answers than expected. One thing that should be kept in mind is that teachers, to some extent, must make the trade off between the practically best method for the students' sake and the practically best method from the teachers' side. This should be done even if the ideal situation is to satisfy the students' preference as much as possible for them to perform well in order to achieve high validity of the test.

2. Investigation of Test Topic by concentrating on interviewers, teachers and raters

The data comes from Nakamura (1996). There are three data sources as follows;

1) Data came from two interviewer-raters through retrospective verbal reports. After the interview test was over, the raters were asked to look back and report on what was happening in their rating during the interview. (See Table 6 for Questionnaire)

2) Data came from a video tape rater through a verbal report. In the process of the videotape evaluation, the rater was asked to speak out about what was happening in his mind as much as possible. (See Table 6 for Questionnaire)

3) Data came from three audio tape raters through retrospective verbal reports. After the tape evaluation was over, the raters were asked to recollect what was happening in the process of rating. (See Table 6 for Questionnaire)

Table 6

1. Questions used for the interviewer-rater
   1) What was happening in your mind when you were interviewing the first student (at the beginning, in the middle and the towards the end of the interview)?
   2) What was happening in your mind when you were interviewing the second student?
   3) Can you tell me who was the most interesting student or the most impressive student?
   4) In the process of one interview (an interview for one person), what were you doing during each of the three stages (at the beginning, in the middle and the towards the end of the interview)?
   5) When you were asking questions or listening to students' responses what were you thinking about?
   6) What was the big difference between what you had anticipated and what you were actually doing in the interview session, in terms of evaluation items or evaluation criteria?
   7) When you changed your questions, or the level of difficulty, did you change your criteria (e.g. from 3 to 2)?
   8) Was it difficult to be an interlocutor and at the same time a rater?
   9) What were you actually thinking about when the student was making a response?
   10) Was there any unexpected finding in/after the interview?

2. Questions used for the tape (audio/video) raters.
   1) What was happening in your mind when you were evaluating the first student?
   2) What was happening in your mind when you were evaluating the second student?
   3) In the process of evaluation, what steps were you going through?
   4) What was the big difference between what you had anticipated before and what happened after the tape evaluation?

5) What was the main problem or the most difficult thing as a rater?
6) Was there any unexpected finding in/after the tape evaluation?

The following research questions were taken into account to get appropriate information.

(1) Is the interviewer a classroom teacher or not?
(2) Can the interviewer-rater share the common topic for conversation with students in the test situation?
(3) Are the raters interested in the topics that the students can easily respond?

Results and Discussion

1) Analysis of the data from interviewer-raters

The content of the questions in the interview is highly dependent on how good the student teacher/interviewer relationship is before the interview. The big difference between the speaking test and the writing test is that the interviewer should have good rapport with the students and create as comfortable an atmosphere as he can before or within the interview session. There is a difference in terms of asking questions between the situation in which an interviewer knows the students well and the situation in which an interviewer knows them very little or not at all.

There is a difference between a situation when an interviewer is a known classroom teacher and a situation when an interviewer is a complete stranger. In a classroom setting, classroom teachers interview the students and grade their speaking ability. In other words, teachers have to think about the context of the questions and also the check sheet. Also, the test date (either at the beginning of the academic year, or in the middle of the academic year, or towards the end of the academic year) should be taken into account because the greeting words or question words would be completely different depending on how well the interviewer knows the students.

In a class interview, the interviewer tries to get the students to speak up. Therefore, the interviewer-rater always thinks about what relevant questions will elicit the students' speaking ability and thus, the interviewer is deeply involved in the content of the students' production as well. This is where the instability of the question-preparation (the inconsistency of the question-presenting), might occur. The reliability might become even lower when different questions are asked to different students. There exists a

dilemma between trying to have students speak up and not being able to keep reasonable reliability.

2) Analysis of the data from a videotape rater through an introspective verbal report

One rater tried to get the context or the situation in which the student was involved by trying to understand what the student was trying to say. He maintains that comprehensibility to the rater is important. He stresses the point that although students should try to make him/herself understood in English, raters should try to understand the students by expanding their knowledge of the topics which students are familiar with. In other words, raters should try to grasp the topics that students are interested in so that raters can share the topic with the students, even in the test situation. Moreover, raters should have as much information as possible about the students' academic background.

3) Analysis of the data from three audio tape raters through a retrospective verbal report

These three raters tried to understand what the students wanted to say. It was not the content but rather the comprehensibility that mattered in the speaking test. The raters' focus for the evaluation was not on grammar or pronunciation, but on total communication, in other words, comprehensibility.

The important thing for the raters is how much background knowledge they can share with students. Raters need to try to increase their knowledge of topics in which students are interested in everyday classroom situations, even though it is sometimes possible to elicit the new information from the students in the test situation. We must take into consideration that the beginning level or intermediate - low level students are more likely to speak out about their familiar topics. This can happen only when the raters elicit the topic and share it with the students. Raters (teachers), although they are sometimes intrigued by the students' mistakes in grammar or pronunciation in the process of evaluation, should keep consistently on the track of comprehensibility by trying to share the common ground related to the topic, chosen by the students. Furthermore, since raters want to share the topic/context with the students through tape by trying to understand what the students are trying to say, they should also expand the topics in the testing situation as well as in the classroom situation so that students can understand them better.

In summary, interview-raters, audio-tape raters, and video tape raters value comprehensibility (to try to understand what the students try to say) highly. In order to make

BEST COPY AVAILABLE

this comprehensibility possible, teachers or raters should expand the topics or the information students are interested in so that they can share them even in the testing situation.

3. Examination of Test Method from the viewpoint of test takers

First we will look at the questionnaire results about the students' preference of test method (whether they like an LL test or they like a face to face interview test).

Table 7

Questionnaire

Which type of a speaking test do you prefer, a test in the language laboratory or a face to face interview test? Why?

Results

LL : Interview =14:30 (n=44)

Comments from the students who are for an LL test

1) I feel tense in the face to face interview test situation.(6)
2) I feel nervous and tense in the face to face test.
3) In the face to face test, it is not fair when different questions are used depending on the level of students.
4) In the LL test I can focus on the sound through the earphones.
5) I feel relaxed when tested with other students in the LL test.(2)

Comments from the students who are for a face to face interview test.

1) I cannot correct sentences once produced in the LL test.
2) There is no human to human relationship in the LL test.
3) I feel less tense in an interview test than in the LL test.
4) The interviewer can wait when I need some time to think.
5) I am not good at handling machine in the LL test.
6) I can take time to think in an interview test. (4)
7) I will be greatly influenced by the neighboring students in the LL test.
8) I am fighting with the time in the LL test.
9) In an interview test the interlocutor reword or paraphrase the unfamiliar words so that I can understand them easily.
10) I don't feel I am speaking in the LL test.
11) It can check our case by case conversation strategies.
12) In an interview test the interviewer would keep the conversation even if I get stuck with some words by offering me some help.
13) The interviewer helps me when I am in trouble in an interview test.
14) I feel tense when I put on an microphone and earphone in the LL test.
15) The LL test is one way test.
16) The interview test seems more practical and real.(4)
17) I can ask when I can not catch the word.

BEST COPY AVAILABLE

Table 7 demonstrates that when we think about face validity and authenticity, the comments from the students who are in favor of a face to face interview test are crucially important. They say that an interview test seems more practical and real. Another comment from them is they feel more cordial atmosphere with the interviewer. Also, they think they can establish much rapport with the interviewer.

However, the comments from the students who are for an LL speaking test should be given another thought especially when the students are nervous about the face to face interview test situations. One comment from this second type should be noticed. The comment says "In the face to face test, it is not fair when different questions are used for different students depending on their level. In an LL speaking test the questions are given equally. This must be fair. So an LL test is much fairer than an interview test." This statement should be taken into account in terms of the fairness issue of testing. Another comment from the LL test side is that they feel tense in an interview test because they are not familiar with native speakers of English. Still another comment is that in the LL test they think they can concentrate on the sounds through the earphones.

Secondly, we will use the data from Nakamura (1996) to explain the factorial structure of the LL speaking test ability and an interview test ability, although in the data the writing ability is additionally included for another reason. Let us look at Table 8.

Table 8

Results of Factor Analysis

Note: Abbreviations are as follows:
Methods……I: interview test, T: tape test,W: writing test
Traits……All: comprehensibility, Cont: content, Dis: discourse, Flue: fluency, Gra: grammar, Incom: interactional competence, Pron: pronunciation, Socom: sociolinguistic competence

	Factor 1	Factor 2	Factor 3
I_ALL	.93807	.04837	.17455
I_CONT	.72045	.16152	.39950
I_DIS	.90654	.10953	.23398
I_FLU	.84378	.20972	.22943
I_GRA	.76120	.16728	.09334
I_INCOM	.74474	.15113	.38222

I_PRON	.73095	.03046	.39340
I_SOCOM	.87389	.12819	.17067
I_VOC	.73233	.06267	.39272
T_ALL	.21794	.12093	.86280
T_CONT	.31485	.12894	.82531
T_DIS	.22268	.11845	.87409
T_FLU	.32271	.17259	.79521
T_GRA	.16685	.24088	.82018
T_INCOM	.36917	.49221	.33502
T_PRON	.38565	.24616	.65663
T_SOCOM	.41704	.38407	.51570
T_VOC	.33258	.17001	.81951
W_ALL	**.18299**	**.91777**	**.03184**
W_CON	.28561	.78377	.27079
W_DIS	.09690	.85977	.26881
W_FLU	.02865	.86566	.26218
W_GRA	.01422	.79431	.12939
W_INCOM	.02145	.85609	.22847
W_SOCOM	.11294	.78781	−.10288
W_VOC	.17243	.81338	.13860

Factor	Eigenvalue	Pct of Var	Cum Pct
1	12.48404	48.0	48.0
2	4.47094	17.2	65.2
3	2.45247	9.4	74.6

Table 8 indicates that three factors can be extracted through the factor analysis and these three factors agree with the three different methods designed for the research. Three factors were named Direct Oral Communication Ability (for Interview Test), Semi-Direct Oral Communication Ability (for LL Speaking Test) and Written Communication Ability.

As far as speaking ability is concerned, the two modes (a face to face interview test and a semi-direct LL speaking test) are rather distinct judging from the factor analysis. However, both are measuring a common area of communication ability to some extent.

Nine traits in Direct Oral Communication Ability contribute to become the fundamental elements of the factor while they still maintain their own characteristics which cannot be replaced by other traits. This idiosyncrasy was made clear through the inter-trait correlation coefficients. The same is true for the nine traits in Semi-Direct Oral Communication Ability. The nine traits function as a factor construct element but they still have their distinctiveness.

BEST COPY AVAILABLE

Writing is quite different from the interview and tape testing in the sense of a test channel. An interview and a tape test require the ability to deal with audio while writing does not. An interview and a tape test are distinct because some students feel nervous about speaking to a machine just mechanically,while other students do not feel easy when they must talk to an interlocutor in the interview session even with their teachers. Thus, the test practice can affect the difference of the two abilities.

Lastly, we will examine the relationship of the four different test methods.

Table 9

Variable	Mean	Std Dev	Range	Minimum	Maximum	Sum	N
GROUP	2.09	.69	2.00	1.00	3.00	67.00	32
INTV.NAT	2.22	.94	3.00	1.00	4.00	71.00	32
LL	2.28	.85	3.00	1.00	4.00	73.00	32
WRITING	2.47	.92	3.00	1.00	4.00	79.00	32

Table 9 shows that among the four different test methods the mean score of the group discussion test was the lowest, and the mean score of the writing test was the highest. One possible interpretation for this result is that many students are not familiar with a discussion type test and that in writing they have enough time to rethink and reorganize their own ideas.

Let us investigate the correlation among the four different test methods in Table 10.

Table 10

Correlation among Four Test Methods

	Group	Intv.Nat	LL	Writing
Group				
Intv.Nat	0.61**			
LL	0.61**	0.65**		
Writing	0.49**	0.66**	0.74**	

n=32

** $p < .01$

One of the results in Table 10 is that the correlation between the LL. test and the writing test is the highest. One possible reason for this is that the test format between the two methods was the same. There were nine contexts and the students were asked to answer

BEST COPY AVAILABLE

the questions and write them down in the writing test, while students were asked to give answers and record them on the tape in the language laboratory. So the only difference was the answering channel.

The correlation between the writing test and the group test is the lowest. One possible explanation is that there is less time to think deeply in the group discussion whereas the writing test allows the students to have enough time to think.

Overall, Table 10 demonstrates that four different methods are relatively interrelated as the measure of the language ability; nevertheless, they are distinct to some extent as indicated in the coefficient. Therefore, each test method should be retained since it is playing its idiosyncratic role to assess the students' language ability from different viewpoints.

VI. Conclusion and Implications

The full study will be left for future research; however, it has been found that different level students perform differently in direct speaking tests and semi-direct speaking tests. It has also been noted that interviewers' choice of test questions influences students' performances and may even affect the raters' ratings.

The fairness issue cannot be framed in absolutes but instead must center on the trade-offs that are sometimes necessary in testing theoretically desirable elements of student production while trying to maintain a relatively high degree of objectivity (Brown 1996). Language teachers should always take into account the possible condition of each institution to conduct a fair judgment of students' language proficiency.

Note

This paper is based on the presentation at the 19th Language Testing Research Colloquium(LTRC) in Orlando, Florida, USA, March 6-9, 1997.

References

Alderson, J.C., C. Clapham & D. Wall. (1995). *Language test construction and evaluation*. Cambridge, UK: Cambridge University Press.

Bachman, L.F. & A.S.Palmer. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Brown, J. D. (1996). *Testing in language programs*. NJ: Prentice Hall Regents.

Brown, H. D. & Gonzo, S. (Eds.). (1995). *Readings on second language acquisition.* Englewood Cliffs, NJ: Prentice Hall Regents.Brown, G.,

Davidson, F. (1996). *Principles of statistical data handling.* London: SAGE Publications.

Genesee, F. & J.A. Upshur. (1996). *Classroom-based evaluation in second language education.* Cambridge; Cambridge University Press.

Haladyna, T. M. (1994). *Developing and validating multiple-choice test items.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Harley, B., Allen, P., Cummins, J. & Swain, M.(Eds.). (1994). *The development of second language proficiency.* Cambridge: Cambridge University Press.

Henning, G. (1987). *A guide to language testing: development, evaluation, research.* Cambridge, MA: Newbury House.

Keeves, J.P. (eds.). *Educational research, methodology, and measurement: an international book.* Oxford: Pergamon Press.

Linn, R. L. (1993). *Educational measurement.* (Third Edition). American Council on Education; ORYX Press.

Lumley, T. (1996). Conflict between fairness and face validity. Paper presented at the 11th World Congress of Applied Linguistics, Jyvaskyla; Finland.

Malmkjaer, K. & Williams, J. (1996). *Performance & competence in second language acquisition.* Cambridge: Cambridge University Press.

McNamara, T. (1996). *Measuring second Language performance.* London: Longman.

Nakamura, Y. (1993). Measurement of Japanese college students' English speaking ability in a classroom setting. Unpublished Ph.D dissertation, International Christian University.

Nakamura, Y. (1996). A study of raters' scoring tendency of speaking ability through verbal report methods and questionnaire analysis. *The Journal of Communication Studies* 5, 3-17. Tokyo Keizai University.

Nakamura, Y. (1996). Development of an English speaking test---establishing validity through multitrait-multimethod (MTMM) analysis. Paper presented at the 11th World Congress of Applied Linguistics, Jyvaskyla; Finland.

Shohamy, E. (1996). Are some methods fairer than others? Paper presented at the 11th World Congress of Applied Linguistics, Jyvaskyla; Finland.

BEST COPY AVAILABLE

BEST COPY AVAILABLE